Statistical Science
2019, Vol. 34, No. 4, 599–603
https://doi.org/10.1214/19-STS748
Main article: https://doi.org/10.1214/18-STS693, https://doi.org/10.1214/18-STS694

Comment: Statistical Inference from a Predictive Perspective

Alessandro Rinaldo, Ryan J. Tibshirani and Larry Wasserman

Abstract. What is the meaning of a regression parameter? Why is this the de facto standard object of interest for statistical inference? These are delicate issues, especially when the model is misspecified. We argue that focusing on predictive quantities may be a desirable alternative.

Key words and phrases: Regression, prediction, variable importance, effect sizes.

1. INTRODUCTION

Professors Buja, Berk, Brown, Kuchibhotla, George, Pitkin, Traskin, Zhao and Zhang have presented a theory of parametric regression that allows one to use the model without requiring that the model be correctly specified. We congratulate these authors for delivering such a masterful treatment of statistical inference for misspecified models. The ideas in these papers are very important and should be required reading for all statisticians. Indeed, essentially all parametric models are misspecified which makes their work extremely relevant.

The papers discuss a number of issues related to interpretation and statistical inference for parameters in misspecified models. The first paper focuses on the linear model and the second generalizes the results considerably. The formalism of a well-specified model is clearly laid out.

In our discussion, we focus on the linear model for simplicity, though similar points hold in general. We take a critical look at the standard regression parameter β, and discuss an alternative predictive view that may be of interest.

Alessandro Rinaldo is Professor, Department of Statistics and Data Science, Carnegie Mellon University, Pittsburgh, Pennsylvania 15213, USA (e-mail: arinaldo@cmu.edu). Ryan J. Tibshirani is Associate Professor, Department of Statistics and Data Science, Carnegie Mellon University, Pittsburgh, Pennsylvania 15213, USA (e-mail: ryantibs@cmu.edu). Larry Wasserman is Professor, Department of Statistics and Data Science, Carnegie Mellon University, Pittsburgh, Pennsylvania 15213, USA (e-mail: larry@cmu.edu).

2. WHY SHOULD WE CARE ABOUT β?

Well-Specified Model

Consider a standard regression setup where we observe independent draws $(X, Y) \sim P$, satisfying a linear model

$$Y = \beta^T X + \epsilon, \tag{1}$$

where $\mathbb{E}[\epsilon] = 0$. Here $X = (X_1, \ldots, X_d) \in \mathbb{R}^d$ and $\beta = (\beta_1, \ldots, \beta_d) \in \mathbb{R}^d$. Even in this idealized setting, we would argue that β is commonly misinterpreted. Many users (and papers and books) implicitly interpret the components of β causally. This interpretation is valid only if every possible confounding variable is included in the model (or if the data are from a randomized experiment).

For example, suppose that Y measures lung damage, while X_1 measures air pollution. In a typical undergraduate class, we might explain the meaning of β_1 as follows: "If we increase air pollution by one unit, and hold all other covariates fixed, then we expect lung damage to increase by β_1 units." This is, at its core, a causal interpretation of the β_1: it describes the effect of an intervention. If the data are observational, there may be many unobserved confounding variables which makes the interventional interpretation of β_1 incorrect.

Most often, the conclusions in a regression analysis are reported in more cautious language but the causal interpretation lurks below the surface. For example, a negative estimate of β_1 in the lung damage and air pollution example would typically lead to some effort to explain why the estimate is negative. No such explanation is required unless one is attempting to interpret

β_1 causally. Unless the experiment is suitably randomized or unless all confounders are measured, the causal interpretation is wrong. In the literature on causal inference, it is not uncommon to use notation such as $\mathbb{E}[Y \mid \text{set } X_1 = x]$ to emphasize a causal effect. Perhaps we should adopt this notation more widely in statistics. It might remind us that the causal slope is

$$\mathbb{E}[Y \mid \text{set } X_1 = x + 1] - \mathbb{E}[Y \mid \text{set } X_1 = x],$$

and thus β_1, which is not equal to the causal slope, must be interpreted with great care, or else, it should no longer the standard object of interest, and an alternative perspective should be taken; more on this in the next section.

These are not new points, and all statisticians are aware of them (at one level or another). But causation is rarely formalized in textbooks on regression. There may be the standard warning that "correlation does not imply causation," but rigorous presentations in terms of counterfacuals, structural equation models, or directed graphs are typically lacking. As a result, many well-trained statisticians do not question what β actually means, even in the idealized case where linearity holds.

Misspecified Model

Let us now turn to the more realistic situation where the linear model does (1) not hold. In this case, the authors propose that β should be interpreted as a functional of P. In particular, we can define

$$\beta = \Sigma^{-1}\alpha,$$

where $\Sigma = \mathbb{E}[XX^T]$ (assumed invertible) and $\alpha = \mathbb{E}[XY]$. The interpretation is that $\ell(x) = \beta^T x$ is the best linear predictor, or in other words, that β minimizes $\mathbb{E}[(Y - b^T X)^2]$ over b.

This definition is mathematically clean and natural. To statisticians, the idea of a best linear predictor is straightforward: it is just a projection, after all. But we should back up and ask: does this define a useful parameter for practitioners? We suspect many are likely to misinterpret β; that is, they are likely to interpret β as if the linear model was correctly specified. This is in addition to the danger of interpreting β causally.

3. WHAT CAN BE DONE WITHOUT A MODEL?

Variable Importance

Throwing the causal interpretation aside, it is still common to interpret β_j as a measure of importance for variable X_j. But if the linear model is wrong, then perhaps we should look for more direct methods. A general model-free approach to this problem is based on assessing conditional independence. For example, to check if X_j is important we can test

$$H_0 : Y \text{ is independent of } X_j \text{ given } X_{-j}, \tag{2}$$

where $X_{-j} \in \mathbb{R}^{d-1}$ denotes the vector containing all covariates except X_j. Some recent methodological breakthroughs have been made in testing such hypotheses, for example, the model-X knockoffs methodology by Candès et al. (2018). This has the potential to be a highly useful tool for practitioners. However, we must note that testing conditional independence is very difficult in a model-lean world, and in fact, was recently shown by Shah and Peters (2018) to be essentially impossible in the absence of any distributional assumptions. In practice, the assumptions needed to derive a useful test of conditional independence are very strong.

Testing conditional independence is one approach to determining variable importance under model misspecification. In addition to variable importance, practitioners are often interested in *variable effect sizes*, which are quantitative (how much does a variable contribute?) rather than binary (does a given variable contribute or not?). It is unclear how a notion of variable effect size can be derived from a test for conditional independence as in (2). We discuss some model-free possibilities below.

Nonparametric Proportion of Variance Explained

Consider the nonparametric proportion of variance explained (or nonparametric R^2) defined by

$$\theta_j = \frac{\mathbb{E}[(Y - \mu_{-j}(X))^2] - \mathbb{E}[(Y - \mu(X))^2]}{\text{Var}(Y)}, \tag{3}$$

where $\mu(X) = \mathbb{E}[Y|X]$ and $\mu_{-j}(X) = \mathbb{E}[Y|X_{-j}]$. On the positive side, the meaning of θ_j is clear, and it is model-free. Unfortunately, estimating the parameter θ_j is difficult. Indeed, this is a semiparametric problem. This was recently studied by Williamson et al. (2017), who use a plug-in estimator with a bias correction based on the first-order influence function. But the resulting inferences rely on very strong assumptions. Moreover, the influence function vanishes under the null hypothesis that $\theta_j = 0$, leading to invalid confidence intervals.

Predictive Proportion of Variance Explained

Now consider a predictive variant of θ_j in (3),

$$\phi_j = \frac{\mathbb{E}[(Y - \hat{\mu}_{-j}(X))^2] - \mathbb{E}[(Y - \hat{\mu}(X))^2]}{\text{Var}(Y)}, \tag{4}$$

where $\hat{\mu}$ and $\hat{\mu}_{-j}$ are estimates of μ and μ_{-j}, respectively. We fit the estimates to training data, using any algorithm: $\hat{\mu}$ is trained by running an algorithm $\mathcal{A}$ on samples (X^i, Y^i), $i = 1, \ldots, n$, and $\hat{\mu}_{-j}$ is trained using the same algorithm $\mathcal{A}$ on the samples $((X_{-j})^i, Y^i)$, $i = 1, \ldots, n$, once we exclude variable X_j. Importantly, while any algorithm $\mathcal{A}$ can be used, it should be emphasized that the parameter ϕ_j and its interpretation are inextricably tied to $\mathcal{A}$: this parameter represents the inflation in prediction error exhibited by the specific algorithm $\mathcal{A}$ after dropping variable X_j, relative to the marginal variance. If $\mathcal{A}$ is a good-predicting algorithm, then this may be an interesting parameter to make inferences on (if $\mathcal{A}$ is poorly-predicting, then it is probably not).

Lei et al. (2018) considered a parameter as in (4), and called it LOCO (leave-one-covariate-out). Some minor differences: Lei et al. (2018) did not consider the normalized version (scaled by $\mathrm{Var}(Y)$), and proposed absolute instead of squared loss. As shown in Lei et al. (2018) (and further studied by Rinaldo, Wasserman and G'Sell, 2019), it is relatively simple to get distribution-free confidence intervals for a parameter like ϕ_j, *provided that we treat* $\hat{\mu}$ *and* $\hat{\mu}_{-j}$ *as fixed, and interpret the expectation as being with respect to a future pair* (X, Y) *only*. In practice, we accomplish this by sample splitting: we use the first half of the sample to fit $\hat{\mu}$ and $\hat{\mu}_{-j}$, and the second half to evaluate test errors and form a confidence interval. Therefore, formally, the parameter is defined conditional on the first half of the data.

In some problem settings (where we expect new data will never arrive, and a model will remain indefinitely fixed after it has been initially fitted), this conditional perspective is reasonable. But in others (where we expect new data will arrive, and a model will be refitted as soon as it does), this is not the right perspective, and instead a *marginal* perspective should be taken: that is, ϕ_j should be defined and interpreted with respect to the randomness both in (X, Y) and in $\hat{\mu}, \hat{\mu}_{-j}$. Markovic, Xia and Taylor (2017) showed a marginal version of the LOCO parameter can be covered using randomization techniques combined with normal approximations; this is interesting progress, but requires assumptions. As far as we can tell, distribution-free inference for the marginal version of ϕ_j is an open problem.

Shapley Effect Sizes

A principled measure for variable effect sizes can be defined using the Shapley value, which originated in game theory (Shapley, 1953). Consider

$$\sigma_j = \sum_S \frac{|S|!(d - |S| - 1)!}{d!}\big[L(S) - L\big(S \cup \{j\}\big)\big], \tag{5}$$

where S varies over all subsets of $\{1, \ldots, d\} \setminus \{j\}$, and L is any real-valued function on sets, for example, $L(S)$ can represent the prediction error using variables in S only. The Shapley value is the unique measure of variable importance satisfying a few simple axioms such as additivity, symmetry, and others. Of course, the Shapley value (5) presents considerable computational and inferential challenges. Advances were recently made in Chen et al. (2017) for structured data.

Conformal Effect Sizes

We finish by describing a measure that relates to ϕ_j in (4), but is a random variable. Define

$$\Delta_j(X, Y) = \big|Y - \hat{\mu}_{-j}(X)\big| - \big|Y - \hat{\mu}(X)\big|. \tag{6}$$

(We use absolute loss and do not scale by $\mathrm{Var}(Y)$ to adhere to the notion in Lei et al., 2018, but this is unimportant.) This is the increase in *prediction loss* at a test point X, Y, using estimates $\hat{\mu}$ and $\hat{\mu}_{-j}$ fit on an independent training set, say (X^i, Y^i), $i = 1, \ldots, n$. We emphasize that $\Delta_j(X, Y)$ is a random quantity. As shown in Lei et al. (2018), we can produce a distribution-free prediction interval for $\Delta_j(X, Y)$, that has finite-sample validity, using conformal prediction.

The idea is that, based on (X^i, Y^i), $i = 1, \ldots, n$, we can produce a conformal prediction band C_n satisfying

$$\mathbb{P}\big(Y \in C_n(X)\big) \geq 1 - \alpha \quad \text{for all distributions } P,$$

where $\alpha \in [0, 1]$ is a desired miscoverage level. The probability is taken over the test point (X, Y) (which is unavailable to us) and the training samples (X^i, Y^i), $i = 1, \ldots, n$ (which are available and used to construct C_n). All that we require is that these samples are i.i.d. (or even more weakly, that they are exchangeable). The construction of conformal bands is described in detail in Vovk, Gammerman and Shafer (2005) and Lei et al. (2018). Now define

$$W_j(x) = \big\{\big|y - \hat{\mu}_{(-j)}(x)\big| - \big|y - \hat{\mu}(x)\big| : y \in C_n(x)\big\}.$$

It follows immediately that

$$\begin{aligned}&\mathbb{P}\big(\Delta_j(X, Y) \in W_j(X) \text{ for all } j = 1, \ldots, d\big)\\ &\quad \geq 1 - \alpha \quad \text{for all } P,\end{aligned} \tag{7}$$

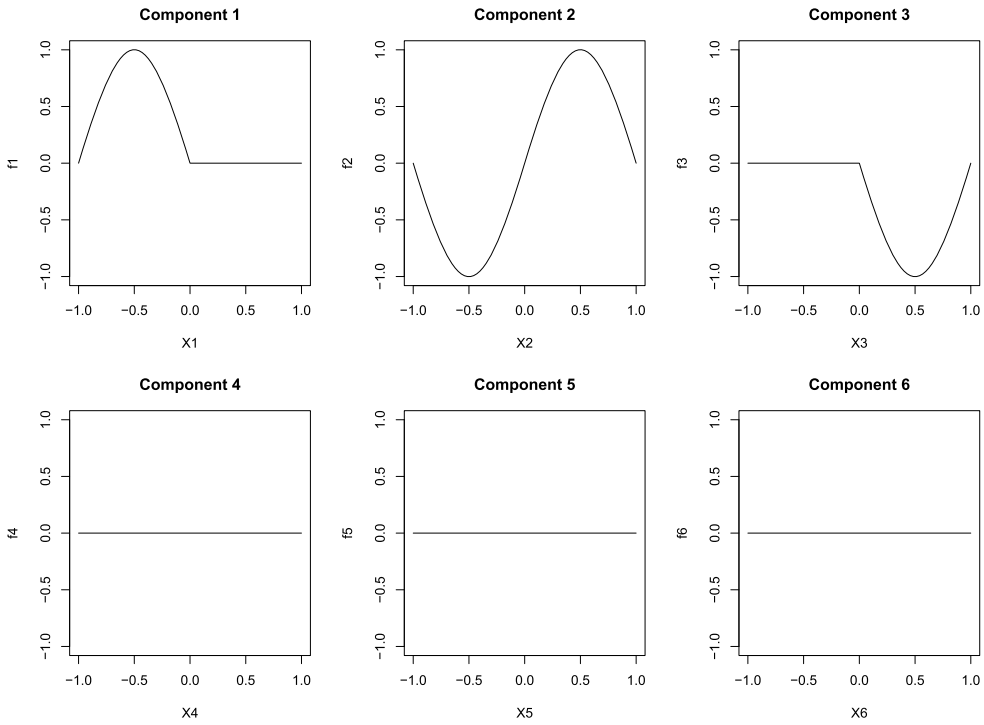

FIG. 1. *Underlying functions in additive models example.*

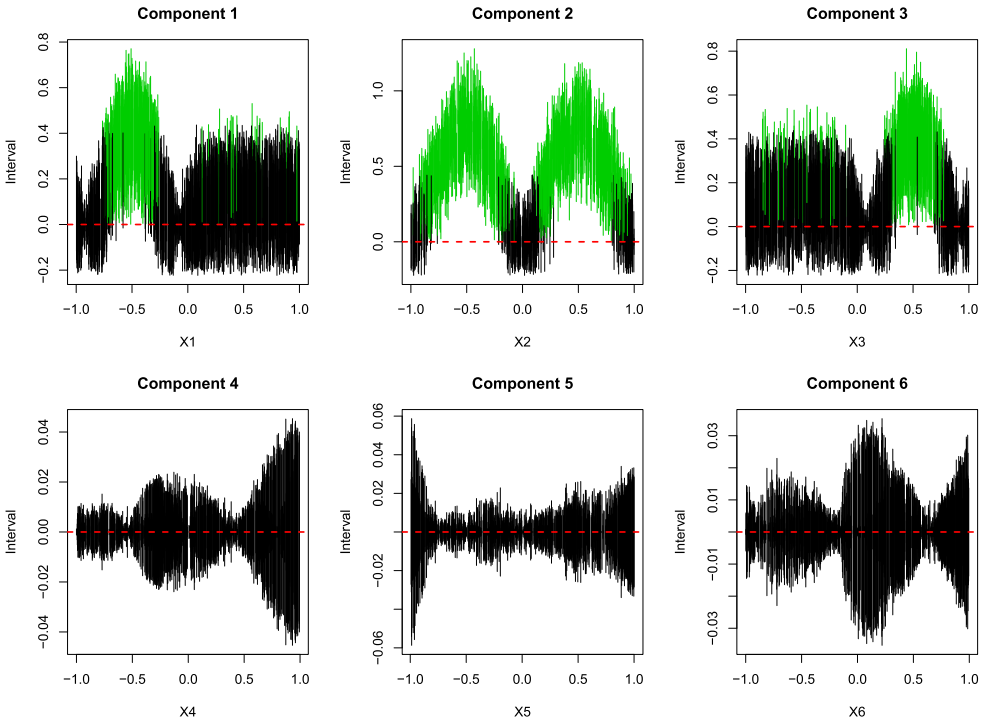

FIG. 2. *Prediction intervals for $\Delta_j(X, Y)$ across all covariates. The prediction intervals are evaluated as X varies over the training covariates. Intervals lying strictly above zero are drawn in green.*

Hence, we have distribution-free, finite-sample coverage for *all variables simultaneously*.

To see how this works in practice, we revisit an example from Lei et al. (2018). The data consist of 1000 samples drawn from an additive model

$$Y = \sum_{j=1}^{6} f_j(X_j) + \epsilon,$$

where $f_4 = f_5 = f_6 = 0$. Figure 1 plots the underlying functions $f_1, \ldots, f_6$. The estimates $\hat{\mu}$ and $\hat{\mu}_{-j}$ were fit using additive regression splines (with each component function having 5 knots) in the definition of (6), and we used conformal prediction in order to obtain intervals with the guarantee (7). Figure 2 plots these intervals as X varies over the training samples. For f_1, f_2, f_3, we see that intervals vary considerably with X, specifically, the intervals which portray a significant increase in prediction loss line up exactly with the regions over which the underlying functions f_1, f_2, f_3 are nonzero.

4. CONCLUSION

The authors have written two compelling papers on statistical inference when models are misspecified. With few exceptions, all models are misspecified. But it is still possible to do inference as long as we interpret parameters correctly. As we have noted in our discussion, there is some room for choosing interpretable parameters. We need not confine ourselves to the standard choices, and in the event of model misspecification, we believe these choices deserve some skepticism, and predictive alternatives may be fruitful.

ACKNOWLEDGMENTS

We thank Rina Barber, Emmanuel Candes, Lucas Janson, Max G'Sell, Jing Lei, Aaditya Ramdas, and Rob Tibshirani for many interesting discussions.

REFERENCES

CANDÈS, E., FAN, Y., JANSON, L. and LV, J. (2018). Panning for gold: 'Model-X' knockoffs for high dimensional controlled variable selection. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **80** 551–577. MR3798878

CHEN, J., SONG, L., WAINWRIGHT, M. J. and JORDAN, M. I. (2017). L-Shapley and c-Shapley: Efficient model interpretation for structured data. Preprint. Available at arXiv:1808.02610.

LEI, J., G'SELL, M., RINALDO, A., TIBSHIRANI, R. J. and WASSERMAN, L. (2018). Distribution-free predictive inference for regression. *J. Amer. Statist. Assoc.* **113** 1094–1111. MR3862342

MARKOVIC, J., XIA, L. and TAYLOR, J. (2017). Unifying approach to selective inference with applications to cross-validation. Preprint. Available at arXiv:1703.06559.

RINALDO, A., WASSERMAN, L. and G'SELL, M. (2019). Bootstrapping and sample splitting for high-dimensional, assumption-lean inference. *Ann. Statist.* **47** 3438–3469. MR4025748

SHAH, R. and PETERS, J. (2018). The hardness of conditional independence testing and the generalised covariance measure. Preprint. Available at arXiv:1804.07203.

SHAPLEY, L. S. (1953). A value for n-person games. In *Contributions to the Theory of Games*, *Vol.* 2. *Annals of Mathematics Studies* **28** 307–317. Princeton Univ. Press, Princeton, NJ. MR0053477

VOVK, V., GAMMERMAN, A. and SHAFER, G. (2005). *Algorithmic Learning in a Random World*. Springer, New York. MR2161220

WILLIAMSON, B., GILBERT, P., SIMON, N. and CARONE, M. (2017). Nonparametric variable importance assessment using machine learning techniques. Univ. Washington Biostatistics Working Paper Series, Working Paper 422.